# Deteksi Video *Deepfake* Menggunakan Ekstraksi Fitur Spasiotemporal

**RINGKASAN DISERTASI**

**Kurniawan Nur Ramadhani**
**NIM: 33219303**
**(Program Studi Doktor Teknik Elektro dan Informatika)**

Institut Teknologi Bandung
Februari 2025

# Deteksi Video *Deepfake* Menggunakan Ekstraksi Fitur Spasiotemporal

Disertasi ini dipertahankan pada Sidang Terbuka Sekolah Pascasarjana sebagai salah satu syarat untuk memperoleh gelar Doktor Institut Teknologi Bandung

Februari 2025

Kurniawan Nur Ramadhani
NIM: 33219303
(Program Studi Doktor Teknik Elektro dan Informatika)

| | | |
|---|---|---|
| Promotor | : | Dr. Ir. Rinaldi Munir, M.T. |
| Ko-Promotor | : | Nugraha Priya Utama, ST, M.A., Ph.D. |

# Deteksi Video *Deepfake* Menggunakan Ekstraksi Ciri Spasiotemporal

Kurniawan Nur Ramadhani
NIM 33219303

## 1. Latar Belakang

Perkembangan teknologi *artificial intelligence*, khususnya dalam bidang *computer vision*, telah memungkinkan terciptanya teknologi yang mampu memanipulasi dan menghasilkan citra atau video secara realistis, termasuk *deepfake*, yaitu konten yang dimanipulasi menggunakan metode berbasis *deep learning*, seperti penukaran wajah dalam video (Mirsky dan Lee, 2021). Meski beberapa *deepfake* digunakan untuk hiburan, seperti mengganti wajah aktor dalam film, banyak yang membawa dampak negatif serius, seperti pada video pornografi dan kampanye hitam. Contohnya, kasus pornografi *deepfake* terjadi saat San Francisco menggugat 16 situs yang menawarkan layanan manipulasi gambar tanpa izin, serta insiden gambar *deepfake* siswa yang tersebar luas di Beverly Hills pada 2024, menyebabkan konsekuensi berat bagi korban. Dalam kampanye politik, *deepfake* sering digunakan untuk memanipulasi video tokoh politik, merusak reputasi mereka. Jumlah video *deepfake* meningkat pesat, dari 7.964 pada 2018 menjadi 14.678 pada 2019, dengan lebih dari 95.000 video aktif pada akhir 2023, 98% di antaranya berupa konten pornografi tanpa persetujuan (Blogs, 2023; Labs, 2019; Spiralytics, 2023). Lonjakan ini menegaskan urgensi pengembangan metode deteksi yang lebih canggih dan adaptif.

Citra dan video *deepfake* memiliki karakteristik unik yang dapat dimanfaatkan sebagai fitur untuk mendeteksi keberadaannya, seperti artefak visual, ketidakkonsistenan geometri wajah, dan pola temporal yang tidak alami. Artefak visual, seperti perbedaan warna mata, bayangan yang tidak wajar, atau kurangnya detail refleksi cahaya, sering muncul akibat keterbatasan metode pembangkitan *deepfake*, terutama dalam menjaga konsistensi detail kecil pada area wajah yang kompleks (Tolosana dkk., 2020). Ketidakkonsistenan geometri wajah, seperti bentuk mata, hidung, atau bibir yang tidak proporsional, disebabkan oleh kesalahan dalam proses pemetaan wajah, terutama ketika data latih tidak mencakup berbagai variasi wajah yang cukup (Yang dkk., 2019). Walaupun saat ini metode pembangkitan *deepfake* telah berhasil membuat konten *deepfake* yang semakin alami, inkonsistensi tersebut masih dapat ditangkap pada tingkat *pixel*. Selain itu, pola temporal dalam video *deepfake*, seperti kedipan mata atau pergerakan bibir yang tidak alami, muncul karena keterbatasan model dalam menghasilkan urutan frame yang konsisten secara temporal (Y. Li dkk., 2018). Fitur-fitur ini mencerminkan ketidaksempurnaan dalam pembangkitan *deepfake*, yang dapat dimanfaatkan untuk mengembangkan algoritma deteksi yang berkinerja lebih tinggi.

Sejak 2017, berbagai metode deteksi *deepfake* telah dikembangkan, dengan fokus utama pada ekstraksi fitur untuk membedakan konten asli dan palsu. Pendekatan

deteksi dapat dibagi menjadi dua kategori utama, yaitu spasial dan temporal. Metode spasial berfokus pada fitur dalam satu citra. Di awal kemunculan *deepfake*, fitur ini masih dapat teramati secara visual, seperti pose kepala, ketidaksinkronan warna mata, serta geometri bagian wajah seperti hidung dan gigi. Akan tetapi efektivitas fitur ini semakin berkurang pada kondisi citra resolusi rendah dan juga berkembangnya kemajuan teknik pembangkitan *deepfake* yang semakin realistis (Y. Li dkk., 2018; Matern dkk., 2019). Selain berbasis visual, ekstraksi fitur spasial juga dapat dilakukan menggunakan metode berbasis fitur lokal, seperti *Image Quality Metric* (IQM). IQM menawarkan kinerja yang lebih baik dibandingkan fitur berbasis visual, tetapi hanya efektif untuk pola sederhana pada tingkat *pixel* dan tidak dapat menangkap hubungan antar *frame* (N. Akhtar dan Dasgupta, 2019). Metode ekstraksi fitur spasial berbasis *deep feature* memanfaatkan metode berbasis *deep learning* untuk mengekstraksi fitur yang lebih kompleks, seperti pada MesoNet, DeepFD, dan *Capsule Forensic*. Namun, beberapa metode ini masih memiliki kelemahan. MesoNet memiliki keterbatasan dalam menangkap pola-pola kompleks pada *deepfake* (Afchar dkk., 2018). DeepFD memiliki ketergantungan pada artefak visual yang mencolok (Hsu dkk., 2018). *Capsule Forensic* dan *MultiTask Learning* mengalami kondisi *overfitting* pada saat menghadapi pola data baru (H. H. Nguyen, Fang, dkk., 2019; H. H. Nguyen, Yamagishi, dkk., 2019a).

Di sisi lain, deteksi *deepfake* berbasis fitur temporal mengandalkan pola perubahan antar *frame* dalam video, seperti yang diterapkan dalam *Recurrent Neural Network* (RNN) dan *Optical Flow*. Meskipun efektif untuk mendeteksi pola temporal, metode ini juga memiliki kelemahan. RNN sering mengalami masalah *vanishing gradient*, terutama pada video yang panjang, sedangkan *Optical Flow* sangat bergantung pada analisis gerakan antar *frame*, sehingga kurang efektif untuk mendeteksi manipulasi yang halus atau yang terjadi hanya pada satu frame (Amerini dkk., 2019; Guera dan Delp, 2018).

Dari metode-metode *deepfake* yang telah dikembangkan pada penelitian sebelumnya, ditemukan beberapa tantangan dalam deteksi *deepfake*. Metode-metode deteksi *deepfake* tersebut memiliki keterbatasan dalam menangani variasi teknik *deepfake* baru dan sering kali sulit untuk melakukan generalisasi pada *dataset* yang bervariasi. Selain itu, metode-metode tersebut memiliki keterbatasan dalam menangkap pola temporal yang kompleks. Kendala lainnya yang ditemui adalah metode deteksi *deepfake* yang ada hanya fokus pada fitur spasial atau temporal secara terpisah, sehingga kurang efektif dalam mendeteksi *deepfake* yang memiliki kombinasi pola manipulasi spasial dan temporal. Kompleksitas yang tinggi juga menjadi kendala bagi metode deteksi *deepfake* dikarenakan berpotensi terjebak pada *overfitting*. Dengan berbagai kelemahan pada masing-masing pendekatan, penggabungan fitur spasial dan temporal menjadi kunci untuk menciptakan sistem deteksi *deepfake* yang dengan kinerja yang tinggi.

## 2. Tujuan dan Batasan Penelitian

Berdasarkan masalah yang telah dipaparkan, maka tujuan penelitian disertasi ini adalah mengembangkan metode deteksi *deepfake* yang menggabungkan pendekatan ekstraksi fitur spasial dan temporal untuk mendeteksi video *deepfake* dengan tingkat akurasi yang tinggi. Tujuan penelitian tersebut dapat dirinci lebih lanjut sebagai berikut.

a. Membangun sistem deteksi video *deepfake* yang menggunakan pendekatan ekstraksi ciri spasiotemporal untuk menghasilkan deteksi *deepfake* berkinerja tinggi.
b. Mencapai akurasi deteksi *deepfake* sebesar minimal 20% lebih tinggi dibandingkan metode *baseline* pada *dataset* Celeb-DF versi 2. Metode *baseline* pada penelitian ini adalah Xception yang memiliki kinerja cukup baik pada *dataset* Celeb-DF versi 2.

Untuk mencapai tujuan tersebut, maka perlu adanya sasaran penelitian agar penelitian tepat sasaran dan mendapatkan hasil sesuai tujuan. Batasan penelitian yang diberikan adalah sebagai berikut:

a. Penelitian ini tidak membahas proses lokalisasi area *deepfake* pada video/citra input. Lokalisasi area *deepfake* memerlukan pendekatan tambahan, seperti segmentasi atau deteksi area manipulasi secara spasial, yang membutuhkan sumber daya komputasi dan waktu penelitian yang lebih besar. Dengan demikian, penelitian ini membatasi diri pada deteksi umum untuk menjaga efisiensi dan tetap pada ruang lingkup yang dapat dikelola dalam waktu dan sumber daya yang tersedia.
b. Penelitian ini tidak membahas bagaimana proses restorasi video/citra *deepfake* ke video/citra asal. Restorasi video atau citra *deepfake* ke bentuk aslinya merupakan proses yang sangat kompleks dan membutuhkan teknik pemulihan data yang sering kali belum tersedia pada *dataset*. Proses ini berbeda dari sekadar deteksi, karena memerlukan rekonstruksi berdasarkan informasi tambahan yang mungkin hilang atau berubah. Pembatasan ini memungkinkan penelitian untuk fokus pada deteksi *deepfake*, tanpa memasuki ranah pemulihan atau pemrosesan ulang, yang memerlukan pendekatan yang berbeda.
c. Penelitian ini tidak membahas deteksi jenis manipulasi *deepfake* yang digunakan pada video/citra input. Identifikasi jenis manipulasi membutuhkan analisis mendalam dengan klasifikasi yang berbeda-beda untuk tiap teknik manipulasi, yang akan memperluas ruang lingkup dan kompleksitas penelitian. Dengan demikian, pembatasan ini dibuat agar penelitian tetap fokus pada deteksi *deepfake* secara keseluruhan, memungkinkan hasil yang lebih cepat dan terukur.
d. Fokus penelitian ini adalah deteksi *deepfake* dengan *Region of Interest* pada area wajah. Sehingga, citra dan video yang dapat diproses oleh metode deteksi *deepfake* yang dibangun pada penelitian ini harus mengandung wajah manusia.

## 3. Metode Penelitian

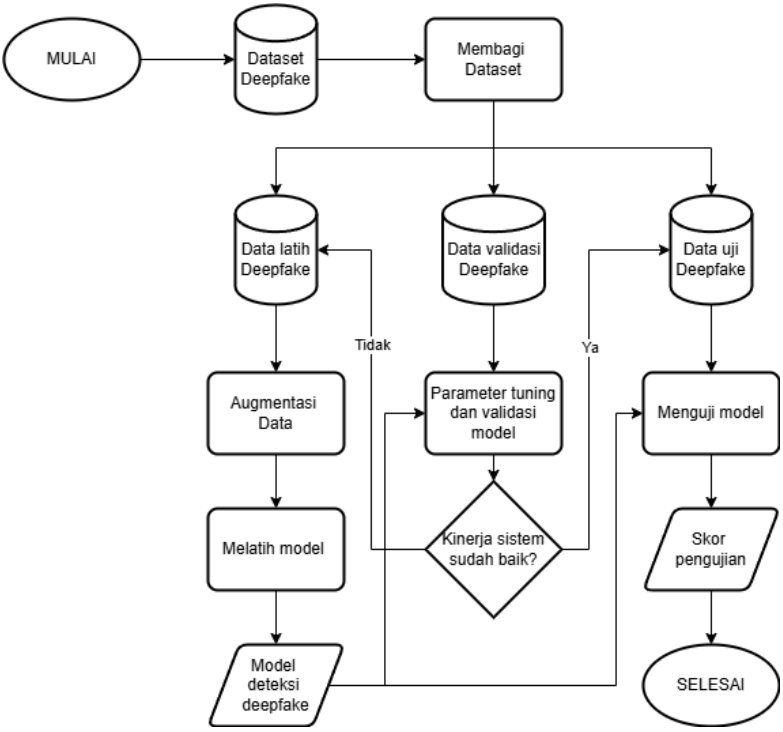
Penelitian disertasi ini mengembangkan metode deteksi *deepfake* yang berbasis pada pendekatan ekstraksi fitur spasial dan temporal. Untuk mencapai tujuan tersebut, penelitian ini dibagi menjadi empat tahapan, yaitu: tahap pertama untuk mengekstraksi area *landmark* wajah, tahap kedua untuk membangun model ekstraksi fitur spasial, tahap ketiga untuk membangun model ekstraksi fitur temporal, dan tahap keempat untuk membangun model deteksi *deepfake* dengan menggabungkan ekstraksi fitur spasial dan temporal menjadi satu model ekstraksi fitur spasiotemporal. Gambar **1** menunjukkan tahapan pelaksanaan penelitian disertasi.


Gambar 1 Tahapan Penelitian Disertasi

a. Tahap 1: tahap untuk mengekstraksi area *landmark* wajah. Pada tahap ini dilakukan penentuan lokasi *landmark* pada wajah, pengambilan area *landmark* pada wajah, dan konstruksi video dari kumpulan area *landmark* tersebut.
b. Tahap 2: tahap untuk membangun model ekstraksi fitur spasial untuk mendeteksi citra *deepfake*. Pada tahap ini dikembangkan metode ekstraksi fitur spasial untuk mendeteksi *deepfake* pada citra. Metode ekstraksi fitur spasial ini berbasis pada arsitektur *Depthwise Separable Convolution* (DSC) dan *Convolution Block Attention Module* (CBAM).
c. Tahap 3: tahap untuk membangun model ekstraksi fitur temporal untuk mendeteksi video *deepfake*. Pada tahap ini dikembangkan metode ekstraksi fitur temporal yang berbasis pada arsitektur *Video Vision Transformer* (ViViT).
d. Tahap 4: tahap untuk membangun model deteksi *deepfake* dengan menggabungkan model ekstraksi fitur spasial dan temporal menjadi satu model ekstraksi fitur spasiotemporal. Model deteksi *deepfake* ini juga dilengkapi dengan *preprocessing* menggunakan ekstraksi area *landmark*.

## 4. Hasil dan Pembahasan

Untuk mendapatkan model deteksi *deepfake* yang baik berdasarkan rancangan arsitektur sistem usulan, dilakukan proses eksperimen dan pengujian model deteksi *deepfake*. Gambar 2 menunjukkan alur eksperimen dan pengujian yang dilakukan pada penelitian ini. *Dataset deepfake* menjadi tiga bagian, yaitu data latih, data validasi dan data uji. Data latih digunakan untuk melatih model deteksi *deepfake* berdasarkan rancangan arsitektur sistem usulan. Data validasi digunakan pada proses validasi model deteksi *deepfake*. Data uji digunakan pada proses pengujian model deteksi *deepfake*.



Gambar 2 Alur eksperimen dan pengujian

Pada proses pelatihan, data latih digunakan untuk melatih model deteksi *deepfake*. Sebelum digunakan, data latih terlebih dahulu diaugmentasi untuk meningkatkan variasi data. Hal ini berguna untuk meningkatkan kemampuan model dalam menggeneralisasi pola data yang dikenali dan menyeimbangkan proporsi setiap label pada data latih. Augmentasi ini dilakukan secara spesifik pada data latih berlabel *real*, karena jumlahnya yang jauh lebih sedikit dibandingkan data berlabel *fake*. Augmentasi yang dilakukan adalah sebagai berikut.

a. Augmentasi spasial yang dilakukan pada setiap *frame*. Teknik augmentasi yang digunakan mencakup transformasi seperti *horizontal flipping*, *rotating*, *add intensity*, dan *multiply*. *Horizontal flipping* dilakukan dengan pertimbangan proporsi citra wajah umumnya bersifat simetris kiri dan kanan. *Rotating* dilakukan dengan mempertimbangkan arah putaran wajah yang rasional yaitu -10 hingga 10 derajat. *Add intensity* dilakukan dengan mengasumsikan bahwa video mungkin mengalami kenaikan intensitas cahaya maksimal 10 *pixel*. *Multiply* merupakan operasi yang mirip dengan *add intensity* namun dengan proses peningkatan intensitas menggunakan faktor pengali, dalam penelitian ini digunakan nilai 0,1.

Dengan teknik ini, model dapat mengenali pola yang lebih luas dan menangkap fitur yang lebih general dari kelas *deepfake*.

b. Augmentasi temporal yang dilakukan dengan memvariasikan frekuensi pencuplikan antara lain 1 *frame* per detik, 2 *frame* per detik, 5 *frame* per detik, dan 10 *frame* per detik.

Model deteksi dilatih menggunakan data latih yang teraugmentasi. Model deteksi yang telah dilatih lalu divalidasi menggunakan data validasi. Skor validasi digunakan untuk menentukan apakah kinerja model deteksi *deepfake* sudah baik atau belum. Jika belum, akan dilakukan proses *hyperparameter tuning* untuk mendapatkan kombinasi *hyperparameter* baru. Kombinasi tersebut digunakan pada proses pelatihan model deteksi *deepfake* untuk mendapatkan model deteksi *deepfake* yang baru hingga didapatkan skor validasi yang baik. Setelah itu, model deteksi *deepfake* diuji menggunakan data uji untuk mendapatkan skor pengujian model. Secara khusus, kegiatan eksperimen yang dilakukan pada penelitian ini adalah sebagai berikut.

a. Eksperimen *hyperparameter tuning* untuk model ekstraksi fitur spasiotemporal pada sistem deteksi *deepfake* usulan. Eksperimen ini bertujuan untuk mendapatkan kombinasi *hyperparameter* model ekstraksi fitur spasiotemporal yang memiliki skor kinerja yang baik. Pada eksperimen ini, *hyperparameter* yang dicobakan adalah *hyperparameter* yang berkaitan dengan arsitektur sistem deteksi *deepfake* dengan mempertimbangkan keterbatasan sumberdaya. *Hyperparameter* yang diujikan dapat dilihat pada tabel berikut. Jumlah *head* merupakan nilai yang menentukan jumlah output satu *layer MultiHeadAttention*. Jumlah *layer* merupakan parameter yang menentukan jumlah *layer Attention* pada *Transformer Encoder* yang digunakan. Nilai dimensi proyeksi menentukan ukuran input vektor pada blok *Transformer Encoder*. Sedangkan parameter *layer* MLP menentukan jumlah *layer* dan struktur *layer* MLP yang ada pada blok *Transformer Encoder*. Tabel 1 menunjukkan nilai yang dicobakan pada setiap parameter.

Tabel 1 Nilai yang dicobakan pada eksperimen *hyperparameter tuning*

| ***Hyperparameter*** | **Nilai yang dicobakan** |
|---|---|
| Jumlah *head* | 8, 16, 32, 64 |
| Jumlah *layer attention* | 4, 8, 16 |
| Dimensi proyeksi | 32, 64, 128 |
| Konfigurasi *layer* MLP | 1 *layer*, 2 *layer* |

b. Eksperimen parameter *preprocessing* sistem deteksi *deepfake* usulan. Eksperimen ini bertujuan untuk mendapatkan kombinasi parameter *preprocessing facial landmark extraction* yang menghasilkan kinerja skor kinerja terbaik. Parameter yang dianalisis pada eksperimen ini adalah sebagai berikut.

Tabel 2 Nilai yang dicobakan pada eksperimen parameter *preprocessing*

| **Parameter** | **Nilai yang dicobakan** |
|---|---|
| Jumlah *landmark* | 16, 25, 36, 49 |
| Ukuran *landmark* | 9x9, 11x11, 13x13 |

c. Eksperimen studi ablasi. Eksperimen ini bertujuan untuk mengamati pengaruh setiap komponen sistem deteksi *deepfake* terhadap kinerja sistem. Komponen yang diujikan adalah modul ekstraksi *landmark*, DSC dan CBAM.

Adapun kegiatan pengujian yang dilakukan pada penelitian ini adalah:

a. Pengujian model deteksi *deepfake* usulan.
b. Pengujian hipotesis untuk membandingkan kinerja model deteksi *deepfake* usulan dengan model *baseline*.

Spesifikasi perangkat yang digunakan pada setiap eksperimen adalah sebagai berikut.

a. *Platform Google Colaboratory*
b. Prosesor dengan kecepatan 2,3 GHz (1 *core*)
c. RAM sebesar 32 GB
d. GPU dengan VRAM 80 GB

Pada setiap eksperimen berlaku *setting hyperparameter* sebagai berikut.

a. *Learning rate* = $1\times10^{-4}$
b. *Epoch* = 100
c. Ukuran *batch* = 8
d. *Optimizer* = Adam
e. *Loss function = categorical cross entropy*

*Learning rate* sebesar $1\times10^{-4}$ dipilih berdasarkan percobaan awal yang menunjukkan konvergensi lebih cepat dan terhindar dari *overfitting*. Jumlah *epoch* 100 digunakan karena untuk *epoch* di atas 100, kinerja sistem cenderung stabil. Ukuran *batch* yang kecil memungkinkan proses pelatihan terhindar dari *local optimum* namun membuat waktu pelatihan menjadi lama karena perhitungan *loss function* dan *update* bobot dilakukan lebih sering. Ukuran *batch* yang besar mempercepat waktu pelatihan namun membutuhkan sumberdaya komputasi yang besar. Dari beberapa kali percobaan, didapatkan ukuran *batch* yang optimal sebesar 8. *Optimizer* Adam dipilih karena mengombinasikan AdaGrad dan RMSprop sehingga dapat menyesuaikan *learning rate* untuk setiap parameter, meningkatkan konvergensi dan kinerja terutama pada data yang *noisy* dan *sparse* (Bock dkk., 2018).

Penelitian ini menggunakan *dataset* Celeb-DF versi 2. *Dataset* ini merupakan hasil pengembangan *dataset* Celeb-DF yang menjadi salah satu *dataset* yang standar digunakan dalam penelitian deteksi *deepfake* (Y. Li dkk., 2020). Tabel 3 menunjukkan proporsi jumlah data yang digunakan pada proses eksperimen dan pengujian. Data latih digunakan pada proses pelatihan model deteksi *deepfake*. Data validasi digunakan pada proses eksperimen berupa *hyperparameter tuning* dan validasi model deteksi *deepfake*. Data uji digunakan pada proses pengujian model

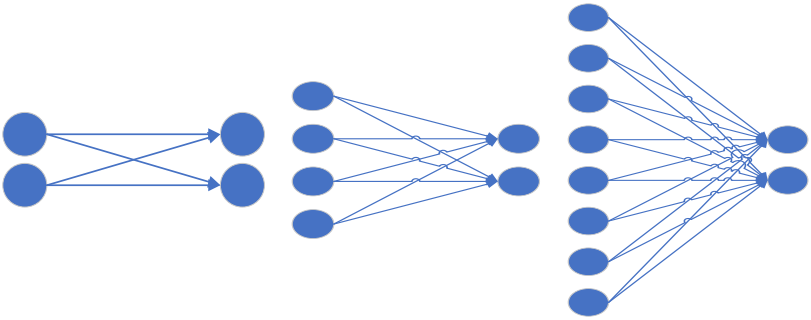
deteksi *deepfake*.

Tabel 3 Proporsi *dataset* eksperimen dan pengujian

| **Kategori** | ***Fake*** | **Asli** | **Jumlah** |
|---|---|---|---|
| Data latih | 6.411 | 1.924 | 8.335 |
| Data validasi | 338 | 52 | 390 |
| Data uji | 984 | 139 | 1.123 |
| **Total** | **7.733** | **2.115** | **9.848** |

Parameter yang diujikan pada eksperimen pertama adalah jumlah *head*, jumlah *attention layer*, dimensi proyeksi, dan *layer* MLP. Pemilihan nilai parameter pada eksperimen ini mempertimbangkan kompleksitas data dan juga ketersediaan sumber daya komputasi. Semakin rendah nilai parameter, semakin sulit dalam memproses data yang kompleks. Semakin tinggi nilai parameter, semakin tinggi kebutuhan komputasi. Pada arsitektur ini terdapat beberapa parameter yang diatur sebagai berikut.

a. Jumlah *head* adalah jumlah *head* pada *layer MultiHeadAttention* yang merupakan inti dari blok *Transformer Encoder*. Pada eksperimen ini, nilai yang dicobakan untuk jumlah head adalah: 8, 16, 32, dan 64.
b. Jumlah *attention layer* adalah jumlah dari *layer Attention* yang ada pada blok *Transformer Encoder*. Pada eksperimen ini, nilai yang dicobakan untuk jumlah *layer* adalah: 4, 8, dan 16.
c. Dimensi proyeksi merupakan ukuran dari vektor input pada *Transformer Encoder*. Pada eksperimen ini, nilai yang dicobakan untuk dimensi proyeksi adalah: 32, 64, dan 128.
d. *layer* MLP yang menentukan jumlah layer dan struktur *layer* MLP yang ada pada blok *Transformer Encoder*. Pada eksperimen ini dicobakan beberapa konfigurasi *layer* MLP, yaitu: 1 *layer* MLP, 2 *layer* MLP dengan *layer* 1 berukuran 1 × dimensi proyeksi, 2 *layer* MLP dengan *layer* 1 berukuran 2 × dimensi proyeksi, dan 2 *layer* MLP dengan *layer* 1 berukuran 4 × dimensi proyeksi. Sebagai ilustrasi dari layer MLP dapat dilihat pada gambar 3 dengan memisalkan dimensi proyeksi=2.



(a) (b) (c)

Gambar 3 Ilustrasi konfigurasi *layer* MLP, (a) *layer* 1=1 × dimensi proyeksi, (b) *layer* 1=2 × dimensi proyeksi, (c) *layer* 1=4 × dimensi proyeksi

Tabel 4 menunjukkan hasil akurasi untuk konfigurasi 1 *layer* MLP. Hasil akurasi tertinggi didapatkan dengan konfigurasi jumlah *head*=16, jumlah *attention layer*=8 dan dimensi proyeksi=32.

Tabel 4 Hasil akurasi untuk konfigurasi 1 *layer* MLP

| **Jumlah *Head*** | **Jumlah *Layer*** | **Dimensi Proyeksi** | **Akurasi** |
|---|---|---|---|
| 8 | 4 | 32 | 72,05 |
| 8 | 4 | 64 | 70,51 |
| 8 | 4 | 128 | 70,51 |
| 8 | 8 | 32 | 72,05 |
| 8 | 8 | 64 | 73,08 |
| 8 | 8 | 128 | 70,77 |
| 8 | 16 | 32 | 70,51 |
| 8 | 16 | 64 | 70,26 |
| 8 | 16 | 128 | 72,82 |
| 16 | 4 | 32 | 71,28 |
| 16 | 4 | 64 | 72,56 |
| 16 | 4 | 128 | 73,08 |
| 16 | 8 | 32 | **74,87** |
| 16 | 8 | 64 | 71,61 |

| Jumlah *Head* | Jumlah *Layer* | Dimensi Proyeksi | Akurasi |
|---|---|---|---|
| 16 | 8 | 128 | 70,77 |
| 16 | 16 | 32 | 74,36 |
| 16 | 16 | 64 | 70,77 |
| 16 | 16 | 128 | 72,05 |
| 32 | 4 | 32 | 71,28 |
| 32 | 4 | 64 | 70,51 |
| 32 | 4 | 128 | 71,79 |
| 32 | 8 | 32 | 70,77 |
| 32 | 8 | 64 | 70,26 |
| 32 | 8 | 128 | 73,59 |
| 32 | 16 | 32 | 72,31 |
| 32 | 16 | 64 | 73,59 |
| 32 | 16 | 128 | 72,05 |
| 64 | 4 | 32 | 73,08 |
| 64 | 4 | 64 | 74,36 |
| 64 | 4 | 128 | 72,56 |
| 64 | 8 | 32 | 71,28 |
| 64 | 8 | 64 | 72,56 |
| 64 | 8 | 128 | 74,36 |
| 64 | 16 | 32 | 72,82 |
| 64 | 16 | 64 | 71,28 |
| 64 | 16 | 128 | 74,1 |

Tabel 5 menunjukkan hasil akurasi dengan konfigurasi 2 *layer* MLP dan *layer* 1 berukuran 1 × dimensi proyeksi. Dengan konfigurasi tersebut, didapatkan hasil akurasi tertinggi sebesar 76,92% dengan jumlah *head*=8, jumlah *attention layer*=4, dan dimensi proyeksi=64.

Tabel 5 Hasil akurasi untuk konfigurasi 2 *layer* MLP dan *layer* 1 berukuran 1 × dimensi proyeksi

| **Jumlah *Head*** | **Jumlah *Layer*** | **Dimensi Proyeksi** | **Akurasi** |
|---|---|---|---|
| 8 | 4 | 32 | 76,67 |
| 8 | 4 | 64 | **76,92** |
| 8 | 4 | 128 | 75,13 |
| 8 | 8 | 32 | 75,38 |
| 8 | 8 | 64 | 76,15 |
| 8 | 8 | 128 | 75,64 |
| 8 | 16 | 32 | 75,9 |
| 8 | 16 | 64 | 76,41 |
| 8 | 16 | 128 | 75,64 |
| 16 | 4 | 32 | 76,15 |
| 16 | 4 | 64 | 76,15 |
| 16 | 4 | 128 | 76,92 |
| 16 | 8 | 32 | 76,15 |
| 16 | 8 | 64 | 75,13 |
| 16 | 8 | 128 | 75,13 |
| 16 | 16 | 32 | 75,64 |
| 16 | 16 | 64 | 76,15 |
| 16 | 16 | 128 | 75,9 |
| 32 | 4 | 32 | 76,92 |
| 32 | 4 | 64 | 75,64 |
| 32 | 4 | 128 | 76,41 |
| 32 | 8 | 32 | 75,38 |
| 32 | 8 | 64 | 75,13 |
| 32 | 8 | 128 | 76,15 |
| 32 | 16 | 32 | 75,13 |
| 32 | 16 | 64 | 75,13 |
| 32 | 16 | 128 | 76,41 |
| 64 | 4 | 32 | 76,15 |

| **Jumlah *Head*** | **Jumlah *Layer*** | **Dimensi Proyeksi** | **Akurasi** |
|---|---|---|---|
| 64 | 4 | 64 | 76,41 |
| 64 | 4 | 128 | 76,92 |
| 64 | 8 | 32 | 75,9 |
| 64 | 8 | 64 | 76,41 |
| 64 | 8 | 128 | 75,13 |
| 64 | 16 | 32 | 76,41 |
| 64 | 16 | 64 | 76,92 |
| 64 | 16 | 128 | 76,92 |

Tabel 6 menunjukkan hasil akurasi untuk konfigurasi 2 *layer* MLP dengan *layer* 1 berukuran 2 × dimensi proyeksi. Hasil akurasi tertinggi didapatkan untuk jumlah *head*=16, jumlah *attention layer*=8, dimensi proyeksi=32 dengan skor akurasi sebesar 77,95%.

Tabel 6 Hasil akurasi untuk konfigurasi 2 *layer* MLP dan *layer* 1 berukuran 2 × dimensi proyeksi

| **Jumlah *Head*** | **Jumlah *Layer*** | **Dimensi Proyeksi** | **Akurasi** |
|---|---|---|---|
| 8 | 4 | 32 | 76,92 |
| 8 | 4 | 64 | 77,44 |
| 8 | 4 | 128 | 77,95 |
| 8 | 8 | 32 | 77,69 |
| 8 | 8 | 64 | 77,69 |
| 8 | 8 | 128 | 77,44 |
| 8 | 16 | 32 | 77,18 |
| 8 | 16 | 64 | 76,15 |
| 8 | 16 | 128 | 76,41 |
| 16 | 4 | 32 | 76,92 |
| 16 | 4 | 64 | 76,15 |
| 16 | 4 | 128 | 77,95 |
| 16 | 8 | 32 | **77,95** |

| Jumlah *Head* | Jumlah *Layer* | Dimensi Proyeksi | Akurasi |
|---|---|---|---|
| 16 | 8 | 64 | 76,15 |
| 16 | 8 | 128 | 77,18 |
| 16 | 16 | 32 | 77,18 |
| 16 | 16 | 64 | 76,15 |
| 16 | 16 | 128 | 76,41 |
| 32 | 4 | 32 | 77,95 |
| 32 | 4 | 64 | 77,44 |
| 32 | 4 | 128 | 76,92 |
| 32 | 8 | 32 | 76,15 |
| 32 | 8 | 64 | 77,69 |
| 32 | 8 | 128 | 76,92 |
| 32 | 16 | 32 | 76,67 |
| 32 | 16 | 64 | 76,67 |
| 32 | 16 | 128 | 77,69 |
| 64 | 4 | 32 | 76,15 |
| 64 | 4 | 64 | 77,44 |
| 64 | 4 | 128 | 76,41 |
| 64 | 8 | 32 | 76,67 |
| 64 | 8 | 64 | 76,15 |
| 64 | 8 | 128 | 76,92 |
| 64 | 16 | 32 | 77,95 |
| 64 | 16 | 64 | 77,95 |
| 64 | 16 | 128 | 77,18 |

Tabel 7 menunjukkan hasil akurasi untuk konfigurasi 2 *layer* MLP dengan *layer* 1 berukuran 4 × dimensi proyeksi. Hasil akurasi tertinggi didapatkan untuk jumlah *head*=32, jumlah *attention layer*=16, dimensi proyeksi=128 dengan skor akurasi sebesar 80,26%.

Tabel 7 Hasil akurasi untuk konfigurasi 2 *layer* MLP dan *layer* 1 berukuran 4 × dimensi proyeksi

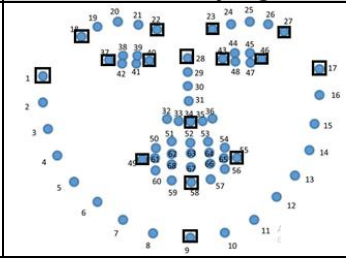
| **Jumlah *Head*** | **Jumlah *Layer*** | **Dimensi Proyeksi** | **Akurasi** |
|---|---|---|---|
| 8 | 4 | 32 | 79,23 |
| 8 | 4 | 64 | 78,21 |
| 8 | 4 | 128 | 77,18 |
| 8 | 8 | 32 | 78,97 |
| 8 | 8 | 64 | 77,69 |
| 8 | 8 | 128 | 78,21 |
| 8 | 16 | 32 | 79,74 |
| 8 | 16 | 64 | 79,23 |
| 8 | 16 | 128 | 77,95 |
| 16 | 4 | 32 | 77,95 |
| 16 | 4 | 64 | 76,92 |
| 16 | 4 | 128 | 78,72 |
| 16 | 8 | 32 | 77,69 |
| 16 | 8 | 64 | 80 |
| 16 | 8 | 128 | 77,95 |
| 16 | 16 | 32 | 80 |
| 16 | 16 | 64 | 79,23 |
| 16 | 16 | 128 | 78,72 |
| 32 | 4 | 32 | 77,18 |
| 32 | 4 | 64 | 77,95 |
| 32 | 4 | 128 | 79,74 |
| 32 | 8 | 32 | 77,69 |
| 32 | 8 | 64 | 80 |
| 32 | 8 | 128 | 77,18 |
| 32 | 16 | 32 | 78,21 |
| 32 | 16 | 64 | 80 |

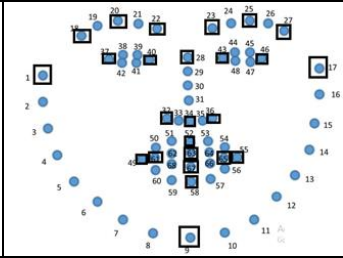
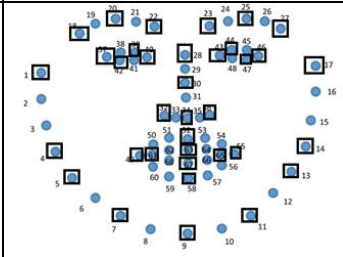
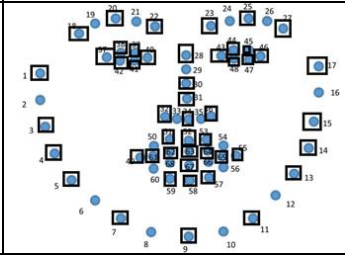
| Jumlah *Head* | Jumlah *Layer* | Dimensi Proyeksi | Akurasi |
|---|---|---|---|
| 32 | 16 | 128 | **80,26** |
| 64 | 4 | 32 | 77,69 |
| 64 | 4 | 64 | 79,23 |
| 64 | 4 | 128 | 78,97 |
| 64 | 8 | 32 | 79,74 |
| 64 | 8 | 64 | 77,69 |
| 64 | 8 | 128 | 77,18 |
| 64 | 16 | 32 | 78,21 |
| 64 | 16 | 64 | 79,49 |
| 64 | 16 | 128 | 80 |

Dengan membandingkan keseluruhan konfigurasi tersebut, didapatkan konfigurasi optimal arsitektur ViViT untuk model deteksi video *deepfake*. Konfigurasi tersebut adalah jumlah *head*=32, jumlah *attention layer*=16, dimensi proyeksi=128 dengan 2 *layer* MLP dan *layer* 1 berukuran 512.

Setelah mendapatkan konfigurasi optimal untuk arsitektur ViViT, dilakukan eksperimen kedua untuk mendapatkan konfigurasi optimal untuk *facial landmark extraction*. Terdapat dua parameter yang diujikan yaitu ukuran *landmark area* dan jumlah *landmark*. Pustaka dlib memungkinkan untuk mengekstrak sebanyak 68 *facial landmark*. Tabel 8 menunjukkan ilustrasi bagaimana pengambilan titik *facial landmark* untuk masing-masing jumlah *landmark*. Tabel 9 menunjukkan hasil eksperimen parameter untuk *facial landmark extraction*. Untuk parameter jumlah *landmark*, nilai yang dicobakan adalah 16, 25, 36, dan 49. Sedangkan untuk parameter ukuran *landmark*, nilai yang dicobakan adalah 9×9, 11×11, dan 13×13.

Tabel 8 Ilustrasi pengambilan titik *facial landmark*

| ***Jumlah Landmark*** | ***Lokasi Landmark yang diekstrak*** |
|---|---|
| *16* |  |

| *Jumlah Landmark* | *Lokasi Landmark yang diekstrak* |
|---|---|
| *25* |  |
| *36* |  |
| *49* |  |

Tabel 9 Hasil eksperimen parameter untuk *facial landmark extraction*

| **Jumlah *landmark*** | **Ukuran *landmark*** | **Ukuran *Tubelet*** | ***Precision*** | ***Recall*** | ***Acc*** | **F1 *Score*** |
|---|---|---|---|---|---|---|
| 16 | 9x9 | 36x36x36 | 0,9672 | 0,784 | 78,97 | 0,866 |
| 16 | 11x11 | 44x44x44 | 0,9674 | 0,79 | 79,49 | 0,8697 |
| 16 | 13x13 | 52x52x52 | 0,9677 | 0,799 | 80,26 | 0,8752 |
| 25 | 9x9 | 45x45x45 | 0,9831 | 0,858 | 86,41 | 0,9163 |
| **25** | **11x11** | **55x55x55** | **0,9865** | **0,864** | **87,18** | **0,9211** |
| 25 | 13x13 | 65x65x65 | 0,9831 | 0,861 | 86,67 | 0,918 |
| 36 | 9x9 | 54x54x54 | 0,9716 | 0,811 | 81,54 | 0,8839 |
| 36 | 11x11 | 66x66x66 | 0,9753 | 0,817 | 82,31 | 0,8889 |
| 36 | 13x13 | 78x78x78 | 0,9754 | 0,823 | 82,82 | 0,8924 |

| **Jumlah *landmark*** | **Ukuran *landmark*** | **Ukuran *Tubelet*** | ***Precision*** | ***Recall*** | ***Acc*** | **F1 *Score*** |
|---|---|---|---|---|---|---|
| 49 | 9x9 | 63x63x63 | 0,9795 | 0,849 | 85,38 | 0,9097 |
| 49 | 11x11 | 77x77x77 | 0,9795 | 0,846 | 85,13 | 0,908 |
| 49 | 13x13 | 91x91x91 | 0,9797 | 0,855 | 85,9 | 0,9131 |

Dari hasil eksperimen tersebut, didapatkan skor akurasi tertinggi adalah 87,18% dengan konfigurasi terbaik untuk *facial landmark extraction* adalah jumlah *landmark* 25 dan ukuran *landmark* 11×11. Skor ini juga menunjukkan adanya peningkatan sekitar 7% jika dibandingkan hanya menggunakan arsitektur ViViT.

Eksperimen berikutnya adalah studi ablasi, yang merupakan studi untuk melihat pengaruh setiap komponen sistem deteksi video *deepfake* terhadap kinerja sistem deteksi tersebut. Sistem deteksi video *deepfake* yang dibangun pada penelitian ini terdiri dari empat modul utama yaitu modul ekstraksi *landmark*, modul DSC, modul CBAM dan modul *Video Vision Transformer*. Modul *Video Vision Transformer* merupakan basis arsitektur sistem sehingga tidak dijadikan sebagai komponen yang diujikan dalam studi ablasi. Sehingga, komponen yang diujikan pada studi ablasi ini adalah modul ekstraksi *landmark*, DSC dan CBAM. Proses studi ablasi dilakukan dengan cara mengubah konfigurasi sistem deteksi *deepfake*. Untuk setiap konfigurasi digunakan kombinasi modul yang berbeda, dengan kombinasi minimal adalah hanya menggunakan *Video Vision Transformer* dan kombinasi maksimal adalah menggunakan keempat modul tersebut. Modul CBAM merupakan pelengkap modul DSC. Maka, penggunaan modul CBAM akan selalu memerlukan penggunaan modul DSC. Tabel 10 menunjukkan hasil studi ablasi yang telah dilakukan. Pada tabel 10, setiap versi merepresentasikan satu jenis kombinasi modul. Versi 1 adalah sistem deteksi *deepfake* yang menggunakan modul *Video Vision Transformer*, DSC dan CBAM tanpa modul ekstraksi *landmark*. Versi 2 adalah sistem deteksi *deepfake* yang menggunakan modul *Video Vision Transformer*, ekstraksi *landmark*, dan DSC tanpa CBAM. Versi 3 adalah sistem deteksi *deepfake* yang menggunakan modul *Video Vision Transformer* dan ekstraksi *landmark*. Versi 5 adalah sistem deteksi *deepfake* yang menggunakan modul *Video Vision Transformer*.

Tabel 10 Hasil studi ablasi

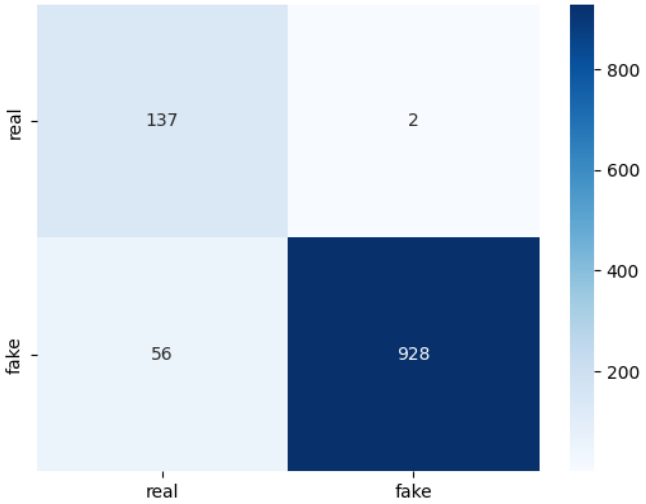
| **Model** | **ViViT** | ***landmark extraction*** | **DSC** | **CBAM** | ***Acc*** | **F1 *Score*** |
|---|---|---|---|---|---|---|
| Versi 1 | √ | | √ | √ | 73,2 | 0,8271 |
| Versi 2 | √ | √ | √ | | 75,69 | 0,8452 |
| Versi 3 | √ | √ | | | 68,83 | 0,7946 |
| Versi 4 | √ | | √ | | 67,76 | 0,7866 |

| Model | ViViT | *landmark extraction* | DSC | CBAM | *Acc* | F1 *Score* |
|---|---|---|---|---|---|---|
| Versi 5 | √ | | | | 52,09 | 0,656 |
| **Model usulan** | √ | √ | √ | √ | **87,18** | **0,9252** |

Berdasarkan tabel 10, dapat disimpulkan beberapa hal. Yang pertama, dengan membandingkan kinerja versi 1 dan model usulan, dapat disimpulkan bahwa penggunaan modul ektraksi *landmark* memberikan efek positif terhadap kinerja sistem deteksi *deepfake*. Hal ini diperkuat dengan membandingkan kinerja versi 2 dan versi 4, serta dengan membandingkan kinerja versi 3 dan versi 5. Proses ekstraksi *landmark* telah berhasil mengurangi kompleksitas dari fitur yang diproses oleh sistem deteksi dengan mereduksi area yang tidak penting pada fitur input. Yang kedua, dengan membandingkan kinerja versi 2 dan versi 3, dapat disimpulkan bahwa penggunaan modul DSC memberikan efek positif terhadap kinerja sistem deteksi *deepfake*. Hal ini diperkuat dengan membandingkan kinerja versi 4 dan versi 5. Penggunaan modul DSC terbukti dapat mengekstraksi informasi penting yang membedakan antara data *fake* dan data asli. Hal ini ditunjukkan juga oleh beberapa penelitian sebelumnya yang menggunakan blok DSC pada arsitektur Xception dan beberapa turunannya. Yang ketiga, dengan membandingkan kinerja versi 2 dan model usulan, dapat disimpulkan bahwa penggunaan modul CBAM juga memberikan efek positif terhadap kinerja sistem deteksi *deepfake*. Hal ini diperkuat dengan membandingkan kinerja versi 1 dan versi 4. Penggunaan modul CBAM memperkaya fitur spasial dari input dengan informasi nilai *attention*. Sehingga, fitur spasiotemporal yang terbentuk memiliki informasi yang lebih lengkap, mencakup nilai *attention* antar *tubelet* dan juga nilai *attention* dari setiap *tubelet*. Dengan menganalisis hasil studi ablasi, dapat disimpulkan bahwa yang paling mempengaruhi kinerja sistem deteksi deepfake adalah kombinasi DSC+CBAM. Hal ini dapat dilihat dengan mengamati akurasi versi 3 yang memberikan selisih dengan akurasi model usulan yang lebih besar jika dibandingkan selisih antara akurasi model usulan dan akurasi versi 1. Hal ini memperkuat argumen bahwa fitur spasial yang dihasilkan oleh kombinasi DSC+CBAM merupakan fitur yang relevan dalam mendeteksi *deepfake*.

Berikut adalah hasil pengujian model deteksi *deepfake* yang didapatkan dari hasil eksperimen. Hasil pengujian diukur menggunakan skor akurasi dan *F1-score*. Gambar 4 menunjukkan *confusion matrix* dari proses pengujian. Dari *confusion matrix* didapatkan nilai akurasi sebesar 94,83% dan *F1-Score* sebesar 0,9697. Tabel 11 menunjukkan contoh data untuk masing-masing hasil deteksi. Kasus *True Positive* adalah video *fake* yang dideteksi sebagai video *fake*. Kasus *True Negative* adalah video asli yang dideteksi sebagai video asli. Kasus *False Positive* adalah video asli yang dideteksi sebagai video *fake*. Kasus *False Negative* adalah video *fake* yang dideteksi sebagai video asli. Variasi contoh data untuk setiap kasus menunjukkan bahwa fitur khas video *deepfake* dan video asli sulit untuk dibedakan dengan pengamatan visual.

Beberapa video *deepfake* yang terlihat alami berhasil dideteksi oleh model deteksi *deepfake* usulan sebagai *deepfake*. Namun ada juga video *deepfake* yang tidak berhasil terdeteksi sebagai deepfake oleh model deteksi *deepfake* usulan.




Gambar 4 *Confusion matrix* hasil pengujian

Tabel 11 Contoh data untuk masing-masing hasil deteksi

| Hasil Deteksi | Contoh Data |
|---|---|
| *True Positive* |  |
| *True Negative* |  |

| Hasil Deteksi | Contoh Data |
|---|---|
| *False Positive* |  |
| *False Negative* |  |

Pada bagian ini dipaparkan tentang pengujian terhadap hipotesis penelitian disertasi ini yaitu: "Model deteksi video *deepfake* yang menggunakan kombinasi modul DSC dan CBAM sebagai ekstraksi fitur spasial dan arsitektur ViViT sebagai ekstraksi fitur temporal serta preproses deteksi *facial landmark* dapat menghasilkan akurasi deteksi video *deepfake* yang tinggi melampaui akurasi metode *baseline* dengan peningkatan akurasi 10% pada dataset Celeb-DF versi 2". Untuk menguji hipotesis tersebut, dilakukan perbandingan antara kinerja model deteksi *deepfake* usulan dengan model deteksi *deepfake baseline*, yaitu model Xception dan model HCiT. Pengujian ini bertujuan untuk menilai apakah model deteksi *deepfake* yang diusulkan mampu memberikan kinerja yang lebih baik secara signifikan dibandingkan Xception dan HCiT. Pengujian dilakukan dengan menggunakan skema *k-fold cross-validation* pada kedua model, dan data akurasi yang diperoleh dianalisis menggunakan *paired t-test*. Hal ini dilakukan untuk menentukan apakah terdapat perbedaan signifikan antara kedua model dalam hal akurasi deteksi *deepfake*. Nilai k yang digunakan sebesar 10. Tabel 12 menunjukkan hasil akurasi dari proses *k-fold cross validation* untuk model deteksi *deepfake* usulan, model Xception, dan model HCiT.

Tabel 12 Perbandingan akurasi dengan metode *baseline*

| **Fold ke-** | **Akurasi (%)** | | |
|---|---|---|---|
| | **Model usulan** | **Xception** | **HCiT** |
| fold-1 | 86,73 | 61,29 | 72,95 |

| Fold ke- | Akurasi (%) | | |
|---|---|---|---|
| | Model usulan | Xception | HCiT |
| fold-2 | 88,05 | 59,47 | 74,56 |
| fold-3 | 87,89 | 60,02 | 71,78 |
| fold-4 | 86,92 | 58,76 | 72,89 |
| fold-5 | 87,34 | 62,11 | 74,24 |
| fold-6 | 88,12 | 59,82 | 73,64 |
| fold-7 | 86,45 | 60,67 | 73,85 |
| fold-8 | 87,56 | 61,9 | 71,72 |
| fold-9 | 87,01 | 58,59 | 73,41 |
| fold-10 | 88,27 | 60,24 | 75,84 |
| rata-rata | 87,43 | 60,29 | 73,48 |
| Varians | 0,41 | 1,47 | 1,41 |

Untuk melakukan uji *t-test*, terlebih dahulu ditentukan hipotesis awal (H0) dan hipotesis alternatif (H1). Hipotesis adalah pernyataan dasar yang mengasumsikan tidak ada perbedaan signifikan antara kelompok yang diuji atau bahwa suatu parameter statistik sama dengan nilai tertentu. Sedangkan hipotesis alternatif adalah pernyataan yang menyatakan bahwa ada perbedaan signifikan antara kelompok yang diuji, atau bahwa suatu parameter statistik berbeda dari nilai tertentu. Dalam pengujian ini, hipotesis awal dan alternatif adalah sebagai berikut:

a. H0: Tidak ada perbedaan signifikan antara akurasi model yang diusulkan dan akurasi model Xception. Dalam konteks ini, H0 menyatakan bahwa akurasi model yang diusulkan sama atau tidak berbeda secara signifikan dari akurasi model Xception.
b. H1: Ada perbedaan signifikan antara akurasi model yang diusulkan dan akurasi model Xception. Dalam hal ini, H1 menyatakan bahwa akurasi model yang diusulkan berbeda secara signifikan dari akurasi model Xception, yang menunjukkan bahwa model yang diusulkan mungkin lebih baik atau lebih buruk.

Hasil perhitungan akurasi menunjukkan nilai rata-rata akurasi dari model yang diusulkan adalah 87,43%, sedangkan rata-rata akurasi dari model Xception adalah 60,29%. Dengan uji *t-test* didapatkan nilai t-statistik dari *paired t-test* adalah 60,7, dengan *p-value* sebesar $4,5 \times 10^{-13}$. Nilai *p-value* yang diperoleh jauh lebih kecil dari tingkat signifikansi standar $\alpha = 0,05$. Hal ini menunjukkan bahwa probabilitas peningkatan akurasi yang diamati terjadi secara kebetulan adalah sangat kecil. Karena *p-value* jauh di bawah tingkat signifikansi standar, H0 yang menyatakan tidak ada perbedaan signifikan antara kedua metode dapat ditolak. Hasil uji *t-test* menunjukkan

bahwa peningkatan akurasi sebesar 27,14% bukan hanya fluktuasi acak, tetapi merupakan hasil dari kontribusi nyata dari kombinasi DSC, CBAM, ViViT, dan *facial landmark detection.* Secara statistik, nilai t-statistik yang jauh dari nol memperkuat argumen bahwa peningkatan tersebut memang bermakna. Konsistensi peningkatan akurasi yang diamati dalam setiap fold dari *k-fold cross-validation* memberikan bukti bahwa model yang diusulkan menghasilkan peningkatan kinerja yang stabil dibandingkan model Xception. Secara keseluruhan, gabungan *p-value* yang rendah, nilai t-statistik yang tinggi, perbedaan akurasi yang besar, dan stabilitas hasil melalui *k-fold cross-validation* membuktikan bahwa model deteksi *deepfake* yang diusulkan lebih unggul dalam akurasi, sehingga hipotesis alternatif (H1) dapat diterima. Dengan hasil ini, metode yang diusulkan terbukti memberikan peningkatan akurasi yang signifikan dibandingkan Xception sebagai model deteksi *deepfake baseline*.

Hal yang serupa dilakukan juga untuk membandingkan kinerja antara metode usulan dengan metode HCiT. Hipotesis awal dan hipotesis alternatif adalah sebagai berikut:

a. H0: Tidak ada perbedaan signifikan antara akurasi model yang diusulkan dan akurasi model HCiT. Dalam konteks ini, H0 menyatakan bahwa akurasi model yang diusulkan sama atau tidak berbeda secara signifikan dari akurasi model HCiT.
b. H1: Ada perbedaan signifikan antara akurasi model yang diusulkan dan akurasi model HCiT. Dalam hal ini, H1 menyatakan bahwa akurasi model yang diusulkan berbeda secara signifikan dari akurasi model HCiT, yang menunjukkan bahwa model yang diusulkan mungkin lebih baik atau lebih buruk.

Dengan uji *t-test* didapatkan nilai t-statistik dari *paired t-test* adalah 35,67, dengan *p-value* sebesar $5{,}27 \times 10^{-11}$. Nilai *p-value* yang diperoleh jauh lebih kecil dari tingkat signifikansi standar $\alpha = 0{,}05$. Hal ini menunjukkan bahwa H0 dapat ditolak dan H1 dapat diterima. Dengan nilai akurasi rata-rata metode usulan lebih tinggi dibandingkan nilai akurasi rata-rata metode HCiT, dapat disimpulkan bahwa metode usulan memiliki kinerja yang lebih baik dalam mendeteksi *deepfake* dibandingkan metode HCiT.

Secara umum, berikut adalah kelebihan dari sistem deteksi *deepfake* usulan dibandingkan dengan metode-metode deteksi *deepfake* lainnya.

a. Efektif dalam ekstraksi fitur spasial dan temporal.

Dengan memanfaatkan *Depthwise Separable Convolution* (DSC) dan *Convolutional Block Attention Module* (CBAM), sistem usulan secara efektif mengekstrak fitur spasial yang penting dengan pemrosesan yang lebih ringan dibandingkan metode berbasis CNN lainnya seperti TD-3DCNN (akurasi 80,52%). Selain itu, penggunaan *Video Vision Transformer* (ViViT) untuk analisis temporal memberikan kemampuan sistem untuk menangkap pola jangka panjang dalam video yang sulit dideteksi dengan metode berbasis *frame* tunggal, seperti EfficientNetB0 (akurasi 88,1%) dan DefakeHop (akurasi 87,65%).

b. Adanya praproses dengan *facial landmark*.

Penerapan *facial landmark* dalam sistem usulan membantu fokus ekstraksi fitur pada area wajah yang relevan, mengurangi *noise* dari latar belakang dan detail yang tidak penting. Hal ini membuat sistem deteksi bekerja lebih efektif dibandingkan model seperti DefakeHop (akurasi 87,65%), yang meskipun ringan, tidak memiliki fokus spasial yang ditingkatkan seperti sistem usulan.

c. Adanya proses augmentasi data.

Proses augmentasi data menjadikan model deteksi *deepfake* dapat beradaptasi dengan kondisi data latih yang tidak seimbang. Augmentasi data latih meningkatkan variasi data sehingga meningkatkan kemampuan model dalam mengenali pola *deepfake* yang variatif.

Adapun kekurangan sistem deteksi *deepfake* usulan dari penelitian ini adalah sebagai berikut.

a. Kesulitan dalam generalisasi pada tipe *deepfake* yang berbeda.

Meskipun sistem usulan menunjukkan akurasi tinggi pada *dataset* Celeb-DF v2, kinerja pada *dataset* dengan variasi yang lebih besar dan kondisi pencahayaan ekstrim belum sepenuhnya teruji. Beberapa metode *state-of-the-art* terbaru, seperti *Self-Attenuated* VGG-16+*Optical Flow* (akurasi 88%) dan EfficientNetB0 (akurasi 88,1%), menunjukkan generalisasi yang baik dalam kondisi lintas domain karena pengaturan arsitektur yang khusus.

b. Kompleksitas arsitektur yang masih cukup tinggi.

Walaupun sistem usulan memiliki kompleksitas yang lebih rendah dibandingkan beberapa metode *state-of-the-art*, arsitektur sistem usulan masih lebih kompleks dibandingkan dengan beberapa model yang lebih sederhana seperti DefakeHop (akurasi 87,65%) dan *Multiscale Spatial Temporal Transformer* (akurasi 87,7%). Kompleksitas ini bisa menjadi kendala dalam penerapan praktis untuk mencapai kinerja optimal.

## 6. Kesimpulan

Pada penelitian disertasi ini, telah dibangun sistem deteksi video *deepfake* yang menggunakan pendekatan ekstraksi fitur spasiotemporal. Pendekatan ini menggabungkan pendekatan ekstraksi fitur spasial dan ekstraksi fitur temporal untuk mendapatkan kinerja deteksi yang tinggi. Untuk ekstraksi fitur spasial, digunakan kombinasi *Depthwise Separable Convolution* (DSC) dan *Convolution Block Attention Module* (CBAM). Untuk ektraksi fitur temporal, digunakan arsitektur *Video Vision Transformer* (ViViT). Selain itu, digunakan juga *facial landmark extraction* sebagai praproses untuk data input yang akan diproses oleh sistem deteksi video *deepfake*. Berdasarkan eksperimen yang telah dilakukan selama penelitian ini, dapat disimpulkan beberapa hal sebagai berikut.

a. Model deteksi video *deepfake* yang berbasis ekstraksi fitur spasiotemporal terbukti berhasil mendeteksi *deepfake* dengan kinerja yang cukup baik. Pada eksperimen, didapatkan skor akurasi sebesar 80,26%. Penggunaan DSC dalam ekstraksi fitur spasial menghasilkan fitur spasial yang relevan dengan deteksi *deepfake*. Relevansi ini ditingkatkan dengan nilai *attention* yang dihasilkan oleh CBAM. Fitur spasial

ini diproses oleh ViViT untuk menghasilkan fitur spasiotemporal yang memiliki relevansi tinggi dalam proses deteksi *deepfake*.

b. Teknik *preprocessing facial landmark extraction* berkontribusi pada peningkatan kinerja model deteksi *deepfake*. Pada eksperimen, *facial landmark extraction* ini berhasil meningkatkan akurasi model deteksi *deepfake* dari 80,26% menjadi 87,18%. Hal ini menunjukkan kontribusi *facial landmark extraction* dalam mereduksi area yang tidak penting pada data *input* sehingga menghasilkan fitur yang tahan terhadap *noise* pada proses ekstraksi fitur spasiotemporal.
c. Model deteksi *deepfake* yang diusulkan menunjukkan keunggulan akurasi yang signifikan dibandingkan dengan model deteksi *deepfake baseline*. Berdasarkan uji hipotesis, model deteksi *deepfake* yang diusulkan terbukti memberikan keunggulan akurasi yang signifikan dibandingkan model deteksi *deepfake baseline*.
d. Hasil pengujian sistem menunjukkan skor akurasi sebesar 94,83%. Skor tersebut membuktikan bahwa fitur spasiotemporal yang dihasilkan dari penggabungan ViViT, DSC dan CBAM berhasil mendeteksi *deepfake*. Dengan skor kinerja yang cukup baik, model deteksi *deepfake* ini berpotensi untuk dikembangkan lebih lanjut dalam implementasi dunia nyata. Model ini dapat diimplementasikan pada *platform* media sosial dan penyedia konten video untuk mendeteksi dan menghapus konten *deepfake* secara otomatis, sehingga membantu mengurangi penyebaran konten yang merugikan. Model deteksi *deepfake* ini juga dapat diimplementasikan oleh lembaga hukum, media, atau perusahaan teknologi untuk melindungi privasi individu dan mencegah penyalahgunaan teknologi *deepfake*.

## 6. Tindak Lanjut

Untuk mengembangkan penelitian ini ke depan, beberapa hal yang disarankan sebagai tindak lanjut adalah sebagai berikut.

a. Sistem deteksi *deepfake* saat ini hanya dapat mengembalikan label asli atau palsu, namun belum dapat mendeteksi teknik pembangkitan citra atau video yang dipakai pada konten palsu. Penelitian ke depan dapat berfokus kepada pendeteksian teknik pembangkitan yang digunakan pada konten palsu.
b. Penelitian sistem deteksi *deepfake* saat ini hanya berfokus pada peningkatan kinerja. Seiring dengan kompleksnya arsitektur yang digunakan, kebutuhan sumber daya juga turut meningkat. Penelitian ke depan dapat berfokus pada efisiensi sumber daya dengan mengefisienkan struktur sistem deteksi namun tidak mengorbankan kinerja.
c. Penelitian sistem deteksi *deepfake* saat ini hanya berfokus pada melabeli konten palsu atau asli. Selain pelabelan, penelitian ke depan dapat berfokus pada restorasi konten palsu untuk mendapatkan konten asli dari konten palsu tersebut.
d. Mengingat keberhasilan augmentasi dalam meningkatkan akurasi deteksi, penelitian lebih lanjut diharapkan dapat mengeksplorasi berbagai teknik augmentasi lain yang mungkin lebih spesifik dan relevan untuk deteksi *deepfake*. Penggunaan augmentasi berbasis temporal juga bisa menjadi tambahan yang

bermanfaat, khususnya untuk data video.

e. Selain augmentasi dan *oversampling*, disarankan untuk mengeksplorasi metode seperti *focal loss* atau *weighted loss* yang dapat membantu model lebih fokus pada kelas yang lebih jarang muncul tanpa memperbesar *dataset*. Teknik ini dapat melengkapi pendekatan augmentasi untuk meningkatkan akurasi lebih lanjut.
f. Dalam pengujian sistem deteksi *deepfake*, variabilitas *dataset* memainkan peran penting. Untuk penelitian lanjutan, disarankan menggunakan *dataset* dengan variasi teknik *deepfake* yang lebih luas, mencakup berbagai jenis manipulasi dan sumber video. Ini akan membantu meningkatkan kemampuan model untuk menggeneralisasi di berbagai situasi nyata.
g. Sebagai saran tambahan, disarankan untuk menguji sistem ini pada *dataset* yang lebih heterogen atau *dataset* nyata dari *platform* media sosial. Hal ini akan memberikan gambaran lebih jelas tentang kinerja sistem deteksi dalam kondisi dunia nyata, yang mana tantangan seperti kualitas video rendah, pencahayaan yang buruk, dan sudut pandang yang berbeda lebih umum terjadi.

**Riwayat Hidup**

Kurniawan Nur Ramadhani lahir di Ujungpandang pada 2 Februari 1988 dari pasangan bapak Jamaluddin (Almarhum) dan ibu Nurmila. Ia meraih gelar sarjana dari Institut Teknologi Telkom (IT Telkom) pada 2008, kemudian melanjutkan studi magister di bidang Informatika di ITB pada 2010 dan lulus pada 2013. Kurniawan menikah dengan Febryanti Sthevanie dan dikaruniai tiga anak: M. Hafidz Nur Rosyad (14 tahun), M. Affan Nur Aulia (12 tahun), dan Hafshah Nurul Aisyah (8 tahun). Sejak 2014, ia menjadi dosen tetap di Universitas Telkom dengan homebase di Program Studi S1 Teknik Informatika.

**Daftar Publikasi Terkait Penelitian**

Ramadhani, K.N., Munir, R., dan Utama, N. P. (2020): A Comparative Study of Deepfake Video Detection Method, International Conference on Information and Communications Technology (ICOIACT), IEEE, hal 394-399. DOI: 10.1109/ICOIACT50329.2020.9331963

Ramadhani, K.N., Munir, R., dan Utama, N. P. (2024): Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attentio, IEEE Access, vol. 2, 8932-8939, https://doi.org/ 10.1109/ACCESS.2024.3352890

**Ucapan Terimakasih**
Rasa terima kasih saya sampaikan kepada Dr. Ir. Rinaldi Munir, M.T., dan Nugraha Priya Utama, S.T., M.A., Ph.D. yang telah memberikan kesempatan, nasehat, bimbingan, pencerahan, dan dorongan terus menerus hingga selesainya disertasi ini.

Ucapan terima kasih disampaikan kepada Prof. Dwi Hendratmo Widyantoro, M.Sc., Ph.D. (Alm), Prof. Dr. Bambang Riyanto, Dr. Masayu L. Khodra, S.T., M.T., Prof. Ir. Hanung Adi Nugroho, S.T., M.Eng., Ph.D. sebagai penguji pada Ujian Kualifikasi, Ujian Proposal, dan Ujian Seminar Kemajuan I - IV serta reviewer buku disertasi, yang telah banyak memberikan arahan dan masukan yang sangat berharga selama proses penelitian disertasi ini dilakukan.

Ucapan terimakasih juga saya sampaikan kepada Prof. Dr. Adiwijaya selaku Rektor Universitas Telkom, Dr. Z.K. Abdurrahman Baizal sebagai Dekan Fakultas Informatika Universitas Telkom, Dr, Arie Ardiyanti Suryani, S.T., M.T. selaku wakil Dekan Fakultas Informatika Universitas Telkom, Dr. Erwin Budi Setiawan, S.Si., M.T. selaku ketua Program Studi S1 Informatika Universitas Telkom, Prof. Dr. Suyanto sebagai kepala Center of Excellence AILO dan Dr. Mahmud Dwi Sulistyo, S.T., M.T. sebagai ketua Kelompok Keahlian Data Science and Intelligent System yang memberikan kesempatan saya untuk melanjutkan studi dan dorongan materil hingga terselesainya doktoral ini.

Penulis juga berterima kasih kepada rekan seperjuangan Program Doktor Teknik Elektro dan Informatika Angkatan 2019 serta rekan Residensi 4: pak Imam, pak Hartanto, bu Ratih, pak Reza, pak Cokorda, pak Luki, pak Idris, serta rekan-rekan lainnya, atas diskusi dan dukungannya yang sangat berharga. Saya ucapkan pula terima kasih yang sebesar-besarnya kepada seluruh pimpinan, rekan dosen, dan staf di Fakultas Informatika Universitas Telkom, khususnya staf dan dosen S1 Informatika.

Rasa terimakasih setinggi-tingginya dipersembahkan kepada Almarhum Bapak, Mama, istri tercinta Febryanti Sthevanie, mertua, adik, dan seluruh keluarga yang telah memberikan semangat dalam menyelesaikan studi S3.